INFORME FINAL:
Prediccion de Engagement en Facebook

Analisis Comparativo, Diagnostico y Modelado Predictivo

Dataset: Cosmetic Brand (Moro et al., 2016)

Generado automaticamente tras analisis exhaustivo en Python

1. Introduccion y Contexto de Negocio

El presente informe detalla el proceso de analisis, limpieza y modelado predictivo realizado sobre un conjunto de datos perteneciente a una reconocida marca de cosmeticos en Facebook. El objetivo principal no es solo predecir metricas, sino entender la dinamica detras de la interaccion del usuario.

A diferencia de enfoques tradicionales, este proyecto distingue dos necesidades de negocio fundamentales:

- Necesidad A (Prediccion a Priori): Estimar el exito de un post ANTES de publicarlo, usando solo variables disponibles al momento de la creacion (Hora, Mes, Tipo).
- Necesidad B (Diagnostico Post-Hoc): Entender las causas del exito UNA VEZ publicado el post, analizando correlaciones con la viralidad (Shares, Reach).

2. Hallazgos del Analisis Exploratorio (EDA)

Se realizo un analisis estadistico riguroso que arrojo los siguientes descubrimientos clave:

- Distribucion del Target: La variable 'Likes' presenta una distribucion log-normal con fuerte sesgo positivo. El test de Shapiro-Wilk confirmo la no-normalidad ($p < 0.05$), justificando la transformacion logaritmica (Log1p) para el modelado.
- Diferencia con la Literatura: Contrario al paper de Moro et al., en este dataset el 'Tipo de Contenido' (Foto vs Video) mostro menor relevancia que las variables temporales y el tamaño de la comunidad (Page Likes).
- Data Leakage: Se detecto una correlacion de Spearman > 0.85 entre 'Shares' y 'Likes', lo que confirma que la viralidad es el predictor mas potente, aunque no este disponible 'a priori'.

3. Definicion de Escenarios de Modelado

Para garantizar la honestidad metodologica, se diseñaron tres escenarios experimentales:

Escenario 1: Paper Original (Benchmark)

Replicacion exacta de las 7 variables usadas por Moro et al. para establecer una linea base.

Escenario 2: Optimizado (Sin Leakage)

Ingenieria de caracteristicas temporal. Se crearon variables como 'Is_Weekend' y 'Time_Segment' (Morning/Evening) para mejorar la prediccion a priori sin usar datos del futuro.

Escenario 3: Lifetime (Diagnostico)

Inclusion de metricas post-publicacion seleccionadas via Information Gain (Mutual Information). Este escenario modela la relacion Viralidad-Exito.

4. Resultados del Benchmark y Evaluacion

Se evaluaron 4 algoritmos (Ridge, RF, XGBoost, SVM) utilizando validacion cruzada de 5 pliegues (K-Fold). La metrica principal fue el MAPE (Error Porcentual Absoluto Medio).

Modelo	Esc 1 (Paper)	Esc 2 (Pred)	Esc 3 (Diag)
Ridge Regression	139.9%	136.2%	>1000% (Fail)
SVM (RBF)	137.8%	157.8%	45.7%
XGBoost	**105.5% (Best)**	105.8%	38.4%
Random Forest	115.0%	114.6%	**36.8% (Best)**

Analisis de Resultados:
- La prediccion a priori (Esc 1 y 2) presenta un error alto (~105%), indicando que el exito en redes sociales tiene un componente estocastico alto que no depende solo de la hora de publicacion.
- El modelo de diagnostico (Esc 3) reduce el error drasticamente al 36% con Random Forest, confirmando que el sistema es predecible si se conoce el alcance (Reach) y la viralidad (Shares).

5. Interpretacion (SHAP) y Conclusiones

El analisis de valores SHAP revelo la 'Caja Negra' de los modelos:
1. En el Escenario 3, la variable 'Shares' domina la prediccion. Un post compartido actua como multiplicador exponencial de Likes.
2. En el Escenario 2, las variables temporales como 'Month' y 'Hour' tienen impacto, pero insuficiente para explicar los outliers virales.

CONCLUSION FINAL:
El proyecto demuestra que es posible diagnosticar con precision el rendimiento (R2 > 0.6) basandose en la viralidad. Sin embargo, la prediccion 'a ciegas' requiere incorporar nuevas fuentes de datos (analisis de imagen/texto) para superar la barrera del 100% de error.